



Data Science  
Research Center  
Amsterdam



# The Data Science Research Center

# The goal

- Become a leading center on data science by developing the new data science discipline

Leveraging our scientific excellence

Leveraging our tools and infrastructure

Reaching out

Educating talents



# Amsterdam research embedding



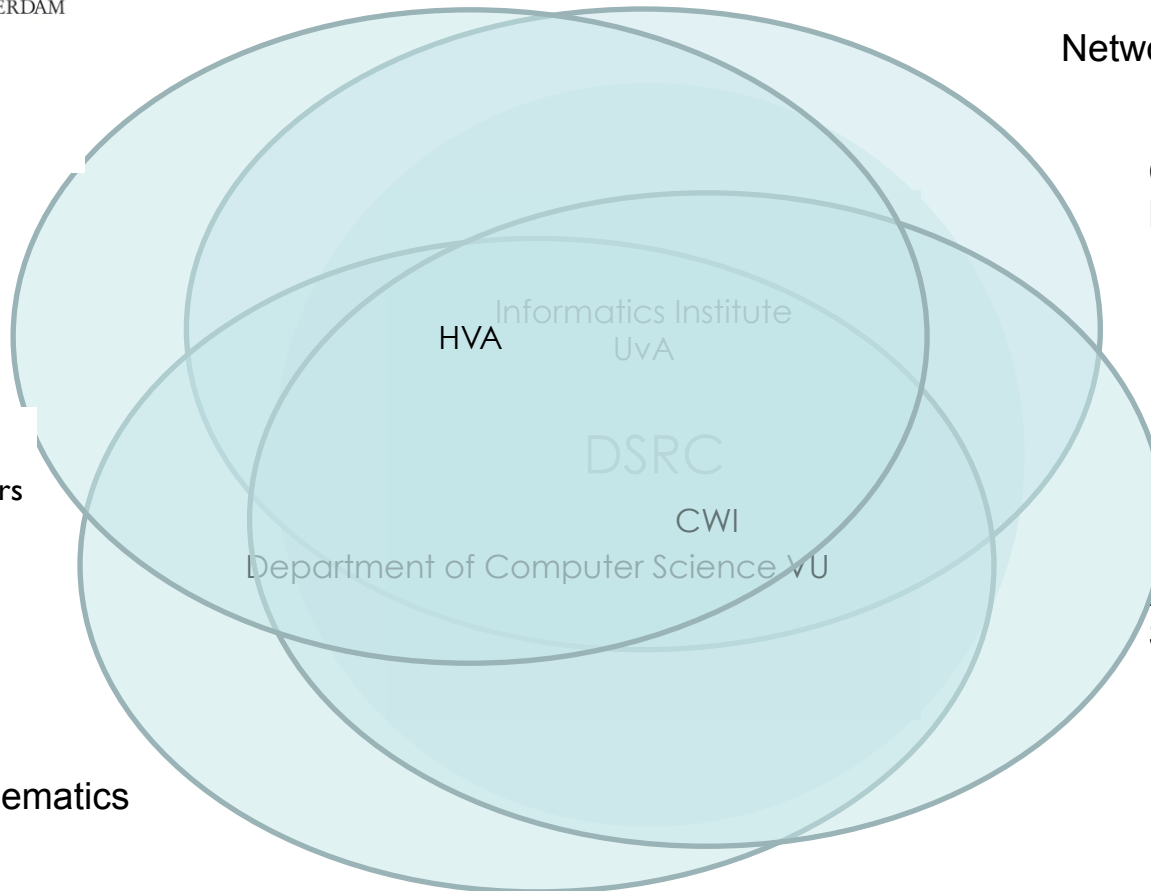
Center for content,  
creation and technology

Network Institute

Center for  
Digital Humanities

CLHC  
Forensic Science

Amsterdam Business  
School



E-science center

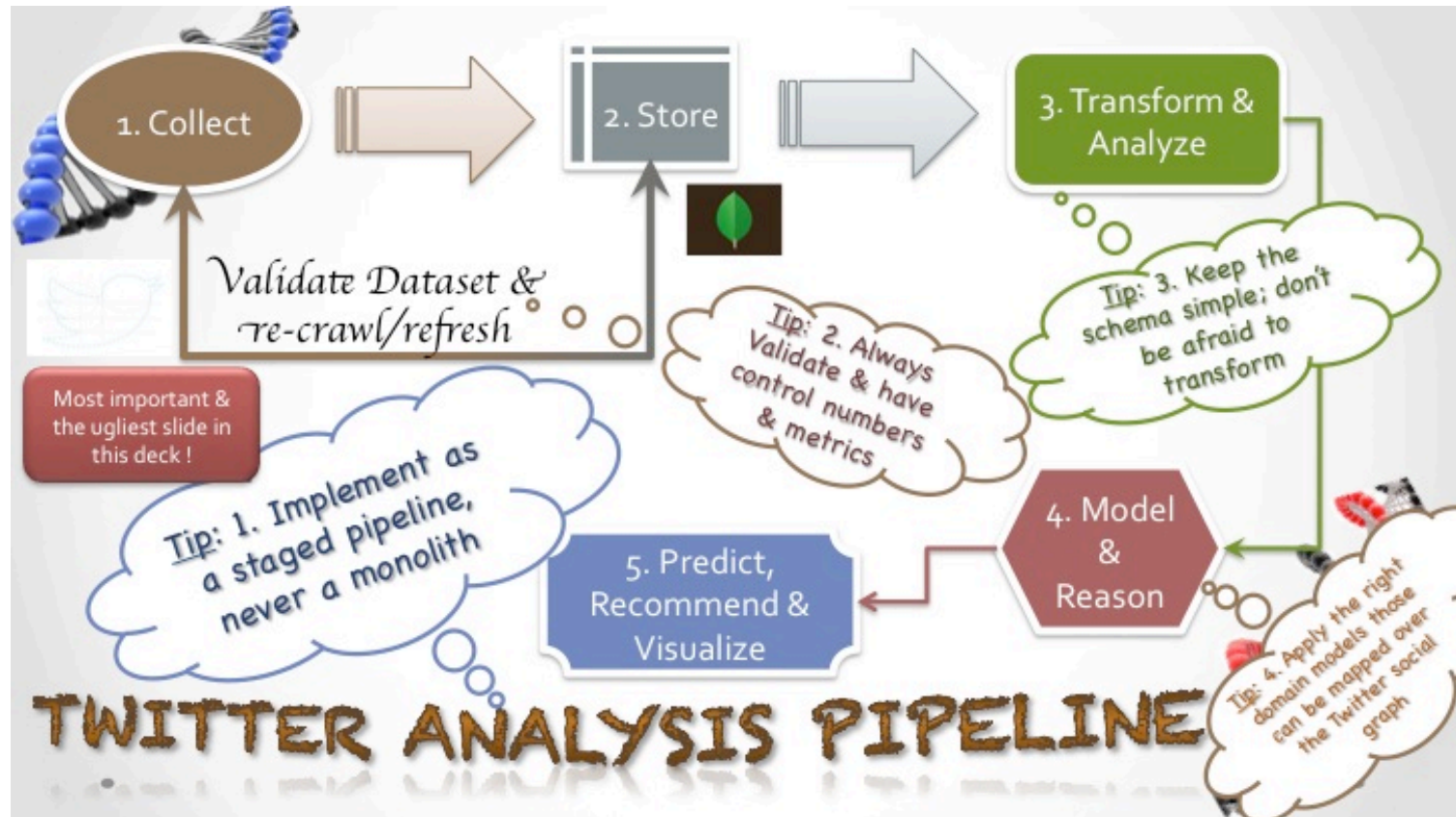
SURFsara

- Founded Fall 2013
- **Four** academic partners
- Built around **multiple** proven research strengths in which we are world leaders

Department of mathematics

ILLC

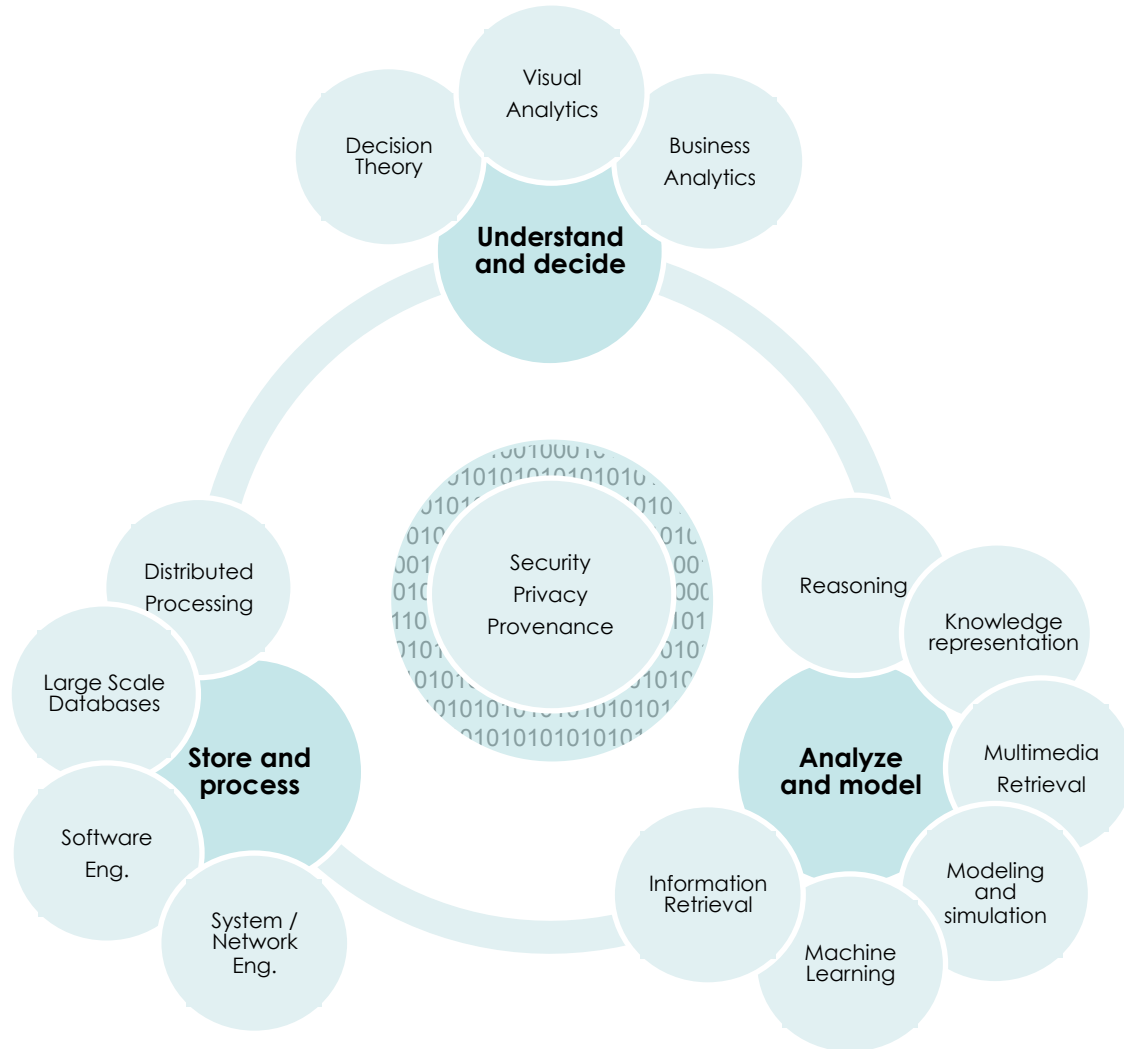
# Big Data Pipeline



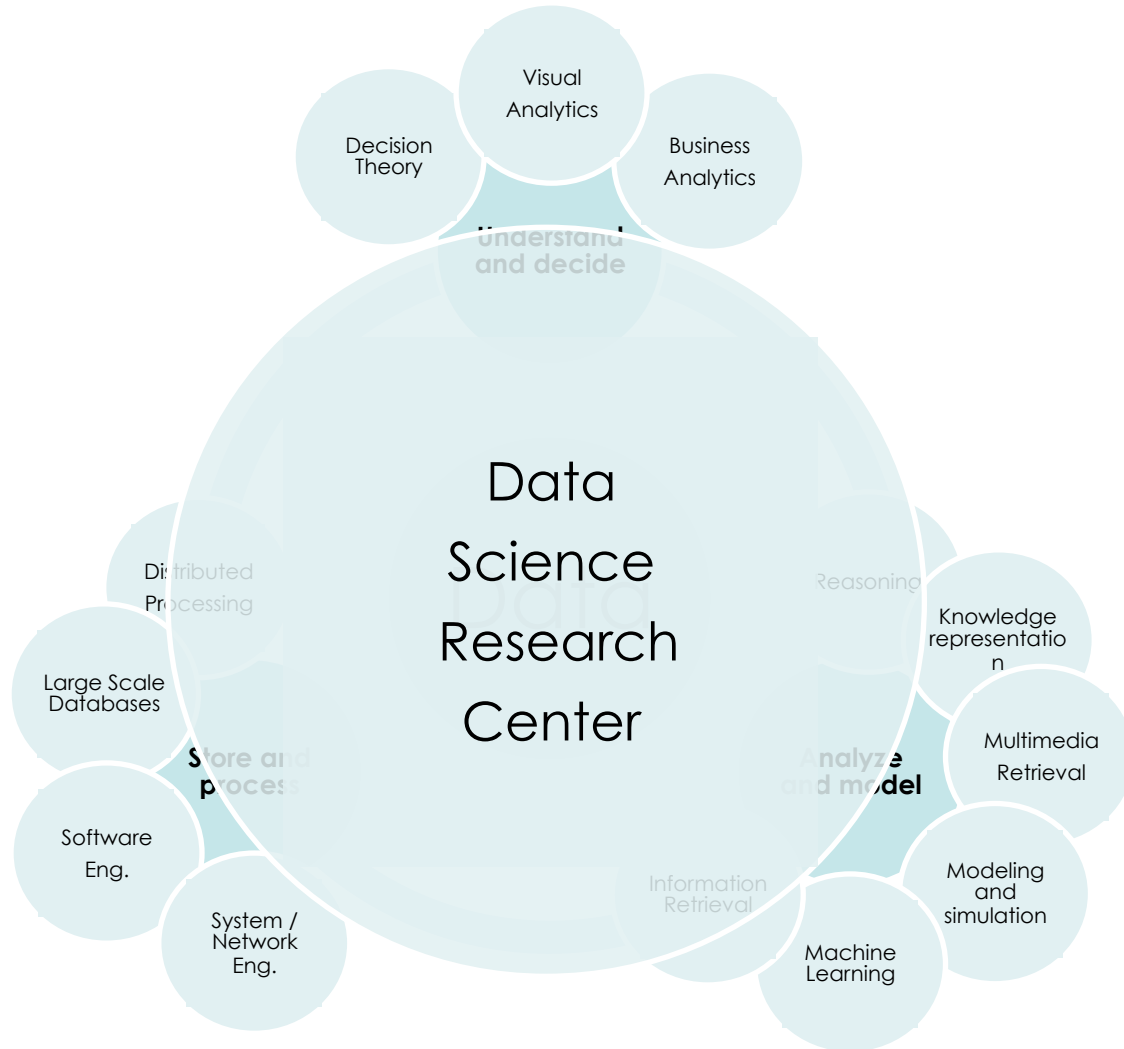
We acknowledge that no single research group has all the necessary expertise.



# Our contributions to data science

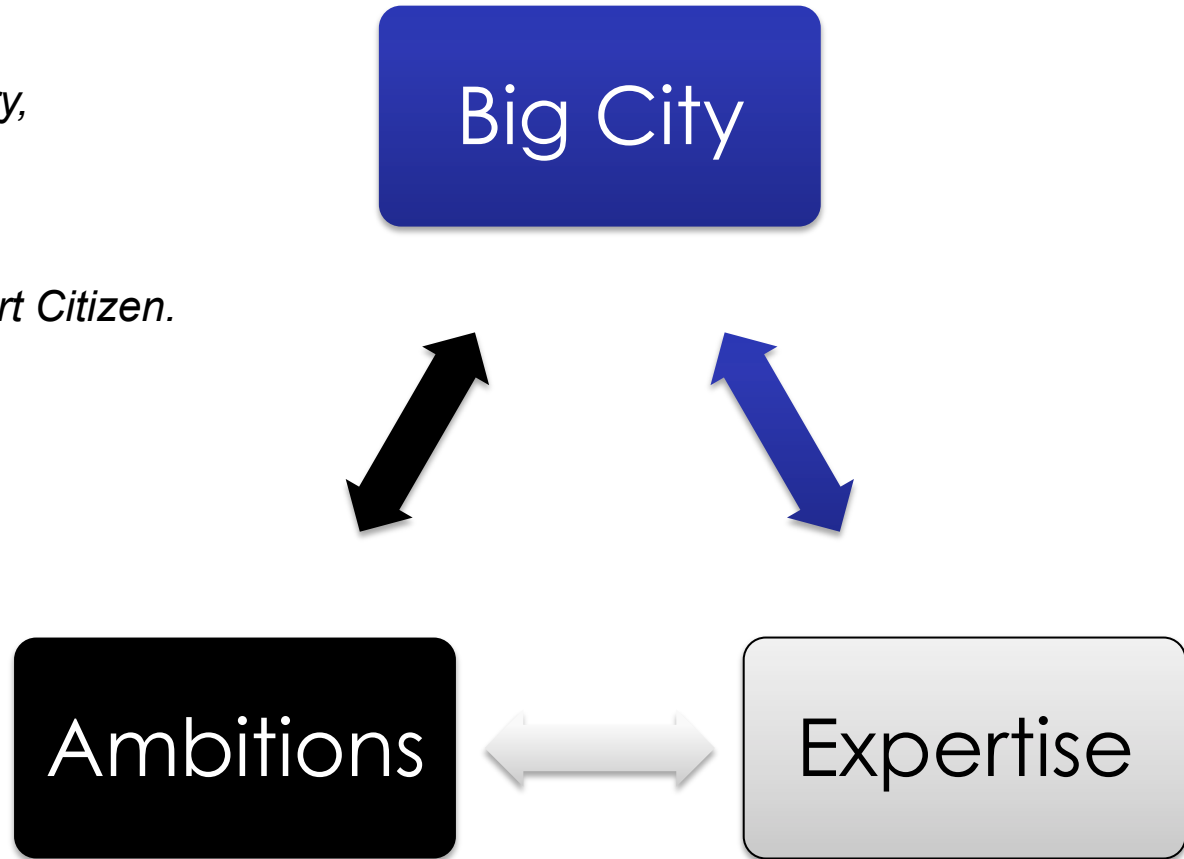


# Research in data science



# The DSRC Trinity

- *Creative Industry,*
- *Health Care,*
- *Finance,*
- *Scientific Data,*
- *Smart City/Smart Citizen.*



- *Quality in academic and applied research,*
- *Focus on new ICT development,*
- *Education,*
- *Entrepreneurship,*
- *Societal impact*

- *Search and analytics,*
- *Machine Learning,*
- *Systems*



# “Big city data” broadly conceived

- **Creative industry**
  - Peter Boncz, Lynda Hardman, Frank van Harmelen, Martin Kersten, Maarten de Rijke, Arnold Smeulders
- **Health care**
  - Ger Koole, Ben Kröse
- **Finance**
  - Gusztai Eiben, Max Welling
- **Scientific data**
  - Henri Bal, Herbert Bos, Maarten van Steen, Cees de Laat
- **Smart city/smart citizen**
  - Frank van Harmelen, Maarten de Rijke

# World class expertise

- **Search and analytics**
  - Peter Boncz, Lynda Hardman, Frank van Harmelen, Martin Kersten, Maarten de Rijke, Arnold Smeulders, Ger Koole, Ben Kröse
- **Machine learning**
  - Guszti Eiben, Max Welling
- **Systems**
  - Henri Bal, Herbert Bos, Maarten van Steen, Cees de Laat



# Ambitions

- First class academic and applied research
  - Focus on new ICT development
- Education
  - BSc, MSc, PDEng
- Entrepreneurship
- Societal impact
- Approach
  - Data-intensive external collaborations
  - Living labs
  - ACE Venture Labs
- International collaborations

# Relation to others (AMS)

- Amsterdam Business School
- Almere Data capital
- Humanities and social sciences
  - Network Institute/ CCCT,
  - IVIR
- Netherlands eScience Center

**I amsterdam.**

NRC  HANDELSBLAD

RIJKS MUSEUM



**Booking.com**

**MARKTPLAATS.NL**

**TOMTOM** 

**ELSEVIER**

amsterdam economic **board**



**waag society**

institute for art, science & technology



# Relation to others (NL)

- Amsterdam  
Metropolitan Solutions  
(TUD, WUR, MIT)
- CHAT: Center for  
Humanities and  
Technology
  - IBM, KNAW, UvA, VU





# Relation to others (EU)

- Academic collaborations across Europe and beyond
  - Cambridge, Oxford, Southampton, Berkeley, Maryland, Tsinghua, Beijing, ...
- Several Networks of Excellence



# Organization

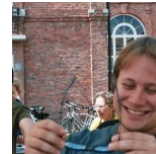


Maarten de Rijke

## Daily management team



Marcel Worring



Paul Groth

Brecht Schipper

## Management board



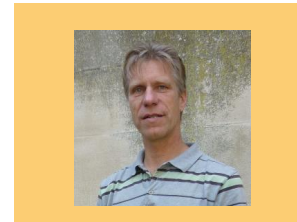
Max Welling



Henri Bal



Bert Bredeweg



Ger Koole

## Leading researchers



Cees de Laat



Sander Klous



Frank van Harmelen



Peter Boncz



Lynda Hardman



Arnold Smeulders



# CWI: Data Management Systems

*Database architecture group at CWI (Boncz, Manegold, Kersten)*



- **Database Systems** research

- DBMS systems design  $\Leftrightarrow$  Computer Architecture
  - Column-store systems, vectorization, open-source, spin-offs and commercial uptake
  - Database algorithms optimizing CPU cache, multi-core
- DBMS systems design for Scientific Data Analysis
  - SCILENS cluster @CWI with high I/O bandwidth
  - MonetDB and R: E.g. real-time detection of short events in LOFAR telescope observations



- **Graph Data** Management

- benchmarks for graph data management systems
  - detecting and exploiting structure in RDF data

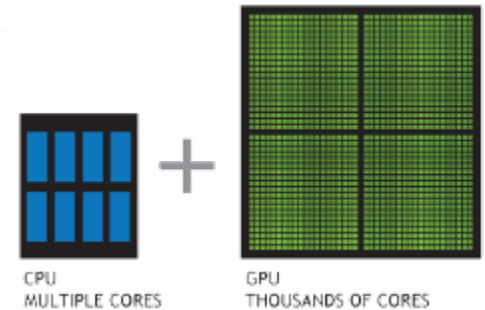


# VU: Glasswing, MapReduce on Accelerators



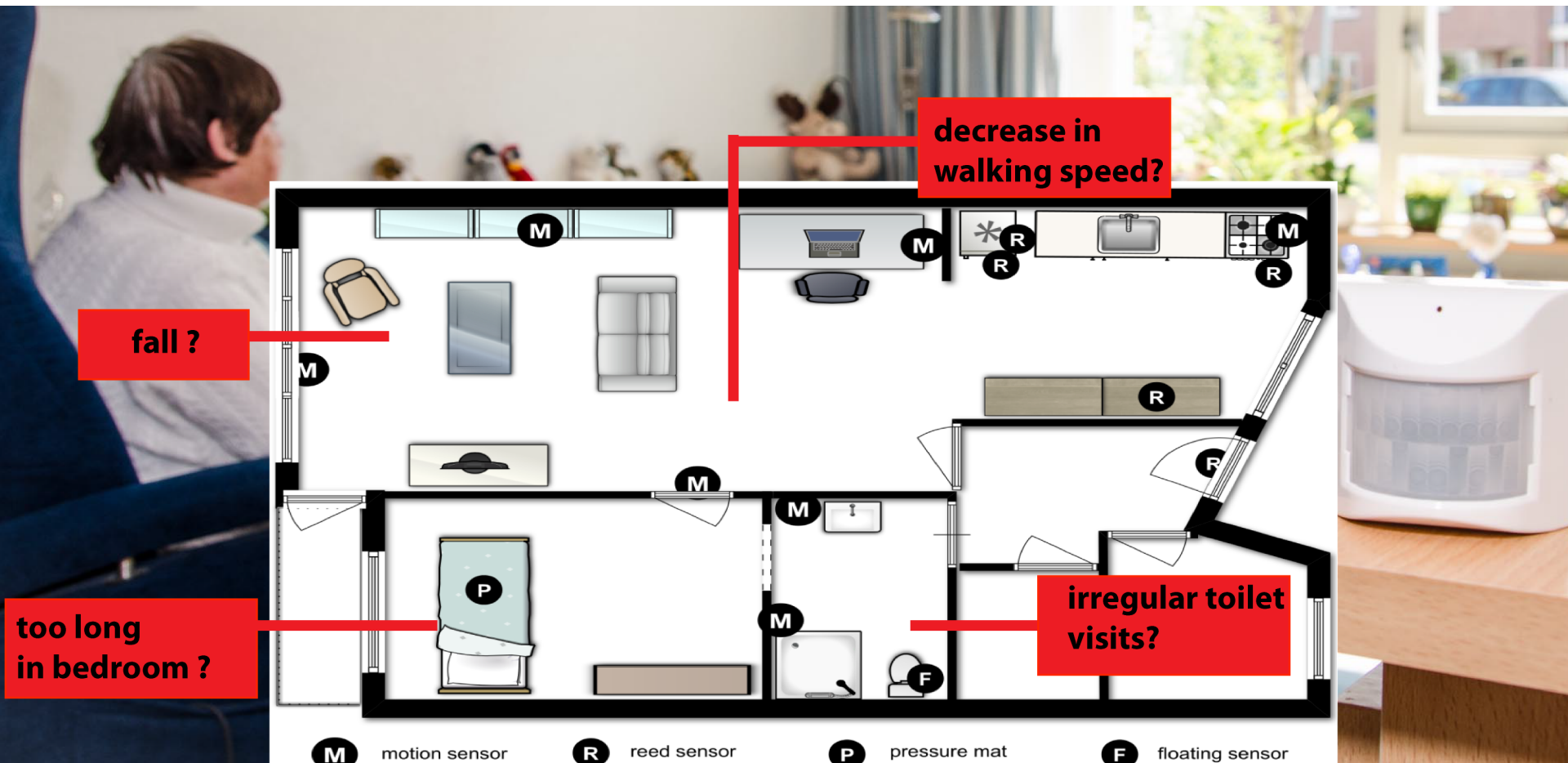
High Performance Distributed Computing Group, H. Bal.

- Runs on a variety of (OpenCL) accelerators
- Massive out-of-core data sets
- Scale vertically & horizontally
- Compute-bound applications
- benefit dramatically from accelerators
- Better scalability than Hadoop



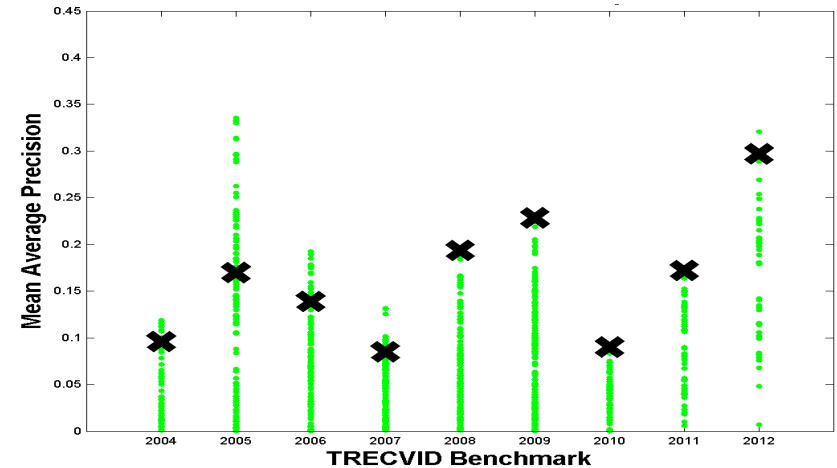
## Home Health Monitoring

- 25 homes equipped with sensors (motion, appliances, bed, doors)
- Over two years data
- Machine learning to detect anomalies, activities, trends in health

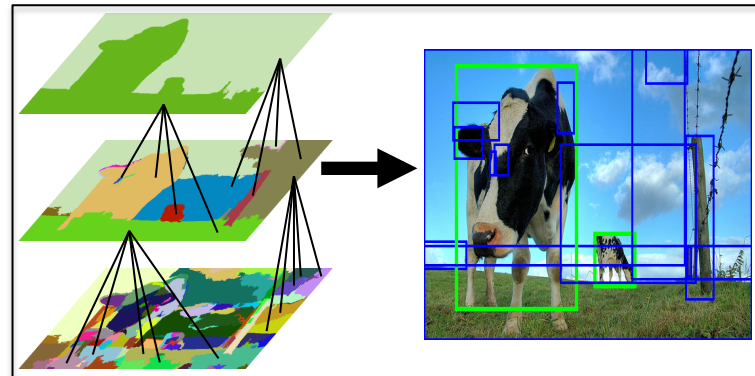
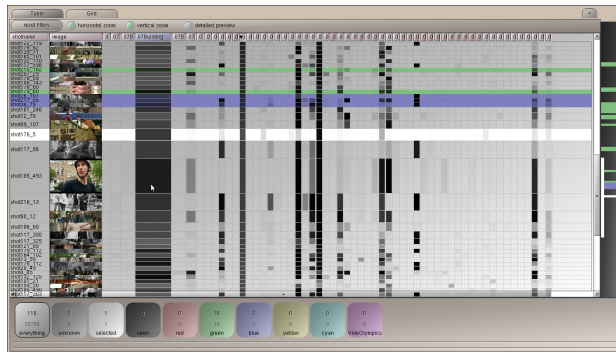


# UVA: Intelligent Sensory Information Systems (ISIS)

"The world is filled with pictures"



We make picture search engines, the best in competition.



And we make systems to analyze large collections of pictures & videos.  
To help the police, social media sites, and consumers.

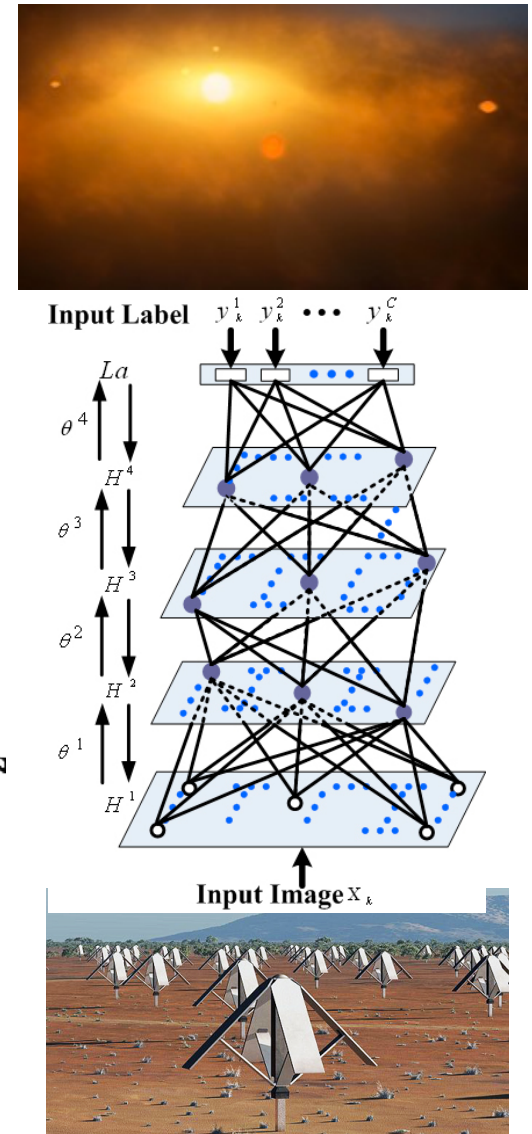


# Machine Learning for Astronomy (IAS)



Square Kilometer Array: 1 exabyte / day in 2024

$$D_{\text{KL}}(Q||P) = \sum_{\mathbf{Z}} Q(\mathbf{Z}) \log \frac{Q(\mathbf{Z})}{P(\mathbf{Z}, \mathbf{X})} + \log P(\mathbf{X}),$$





# Realization

**Research:** a *platform* for research in data science connecting people and methodologies.

**Infrastructure:** a data-driven infrastructure for experimenting with realistic complex data sets.

**Valorization:** a channel between scientific research and third party applications.

**Education:** data-science curricula with realistic data experimentation throughout the program.

# Research: Seed Projects

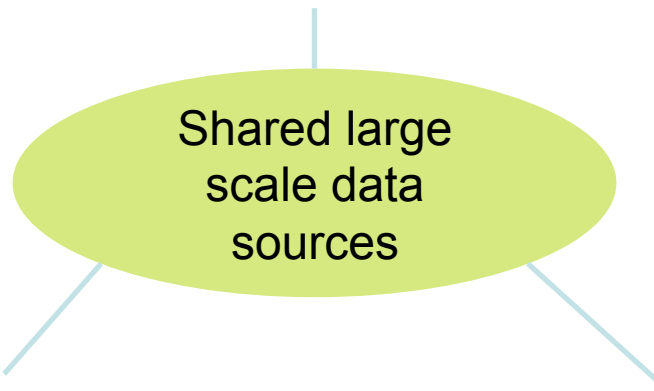
- Accelerated processing of spatio-temporal social graphs
  - Claudio Martella (VU) / Ana Varbanescu (UvA)
- Using Linked Open Data for Medical Question Answering using a combined KR/IR approach.
  - Frank van Harmelen, Annette van Teije (VU), Maarten de Rijke (UvA)
- Data Science in the Browser: Distributed Computing and Visualization of Machine Learning
  - Ted Meeds (UvA) / Magiel Bruntink (UvA)
- Lighthouse: lighting up the warehouse with a SPARQL
  - Spyros Voulgaris (VU) / Peter Boncz (VU/CWI)
- Quantifying Historical Perspectives on WWII
  - Laura Hollink, Victor de Boer (VU) / Jacco van Ossebruggen (VU/CWI), Daan Odijk (UvA)
- BigData2Nets
  - Paola Grosso (UvA) / Patricia Lago (VU)

# Infrastructure

“In a sense, the physical and technical infrastructure becomes invisible and the data themselves become the infrastructure – a valuable asset, on which science, technology, the economy and society can advance.”

[“Riding the wave” EU High Level Expert Group]

Shared domain driven tasks



Shared large  
scale data  
sources

Transparent access  
to distributed computing  
infrastructure

Common tools and  
code bases

# Valorization

- Joint full projects
  - Within the DSRC
  - With industry / governmental organizations
- Small-scale projects
  - From data and problem to solution with quick turnaround
- Competitions
  - From data and problem to innovative solutions worked on by a number of teams
- Spin-offs and startups

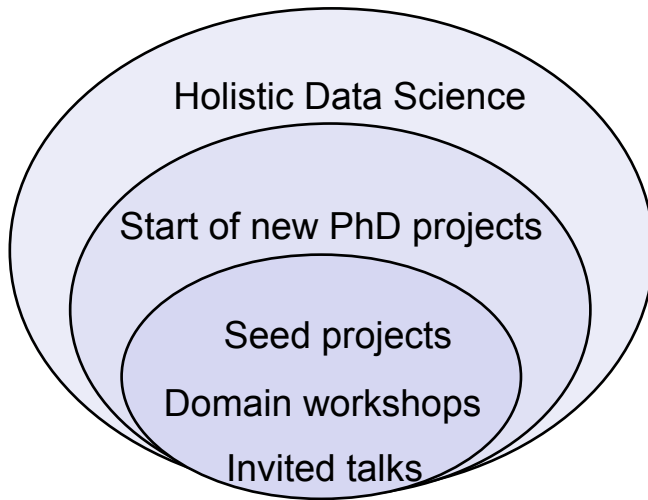


# Education

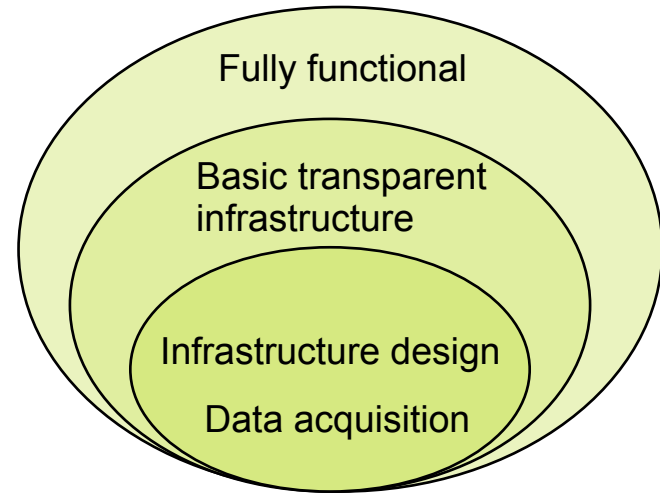
- Data Driven Infrastructure as platform for education in
  - Informatics
    - Information Science, Artificial Intelligence, Software Engineering, Computer Science, Business Analytics
  - Domain specific courses
    - E.g. Minor Data Science for X (your favorite discipline)
  - Commercial courses
- The objective of DSRC
  - to introduce a full data science program with hands-on experience on real data and real problems or innovations



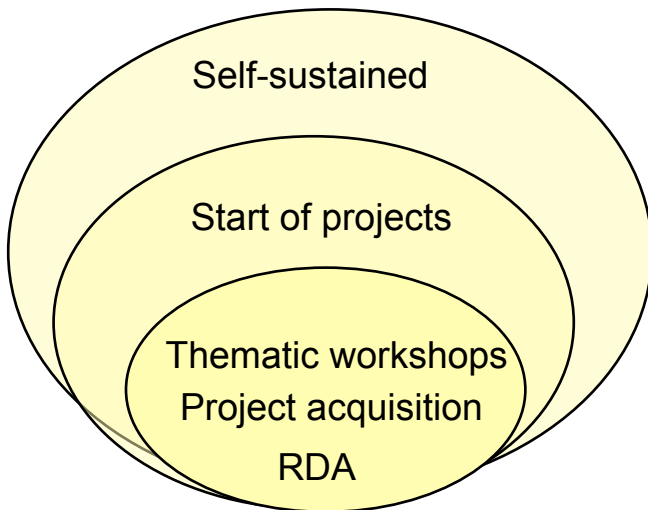
# Roadmap



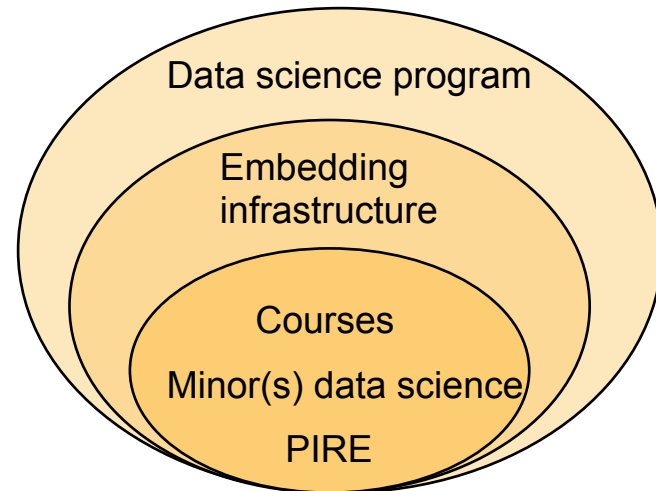
*Research*



*Infrastructure*



*Valorisation*



*Education*

